PR #27396 完整报告

sgl-project/sglang

Cookbook for QAT

合并时间：2026-06-06 02:18

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/27396

执行摘要

本 PR 为 Gemma 4 部署 cookbook 的交互式命令生成器新增了 Checkpoint 选项，让用户可以在 Standard (BF16) 和 QAT (q4_0-unquantized) 检查点之间切换。选择 QAT 时，组件会自动调整模型路径后缀，同时保持 TP 和内存配置不变。这是一个纯文档 /UX 增强，不涉及任何后端逻辑变更，风险极低。

功能与动机

Gemma 4 系列模型发布了 QAT (Quantization-Aware Training) 检查点版本，用户需要一种直观的方式来生成正确的部署命令。PR 文档模板表明这是一次文档更新，旨在帮助用户轻松切换检查点类型，降低使用门槛。

实现拆解

1. 新增 Checkpoint 选择器(docs_new/src/snippets/autoregressive/gemma4-deployment.jsx) - 在选项配置对象中新增 checkpoint 字段，包含 standard（默认）和 qat 两个选项。 - qat 选项的 subtitle 为 q4_0-unquantized，提示用户该检查点的量化类型。
2. 调整模型路径生成逻辑（同上文件） - 在 generateCommand 函数中，根据 values.checkpoint 的值动态计算 qatSuffix：选择 QAT 时追加 -qat-q4_0-unquantized，否则为空。 - 将模型路径变量从 modelNames[modelSize] 改为 ${modelNames[modelSize]}${qatSuffix}，并用于所有引用该路径的位置，包括 --model-path 和 speculative decoding 的 --speculative-draft-model-path。 - 添加注释解释：QAT 发布版本仍使用 BF16 权重，因此 TP 和内存要求与标准检查点一致。
3. 更新文档说明(docs_new/cookbook/autoregressive/Google/Gemma4.mdx) - 在“注意事项”列表末尾新增一行，描述QAT检查点的可用性，并强调其与标准检查点在硬件需求上兼容。

docs_new/src/snippets/autoregressive/gemma4-deployment.jsx

核心变更文件：在交互式部署命令生成器中新增了 Checkpoint 选择器，并调整了模型路径拼接逻辑以支持 QAT 后缀。

```
// docs_new/src/snippets/autoregressive/gemma4-deployment.jsx // 新增 checkpoint 字段，让用户选择 Standard (BF16) 或 QAT 检查点 exportconstGemma4Deployment=()=>{ constoptions={ modelSize:{ name:'modelSize', title:'Model Variant', items:[ {id:'e2b',label:'E2B (~2B)',default:false}, {id:'e4b',label:'E4B (~4B)',default:true}, {id:'12b',label:'12B (Dense)',default:false}, {id:'31b',label:'31B (Dense)',default:false}, {id:'26b-a4b',label:'26B-A4B (MoE)',default:false}, ] }, // --- 新增：Checkpoint 选择 ---
```

checkpoint:{ name:'checkpoint', title:'Checkpoint', items:[{id:'standard',label:'Standard',subtitle:'BF16',default:true}, {id:'qat',label:'QAT',subtitle:'q4_0-unquantized',default:false},] }, // ... hardware, reasoning, toolcall, speculative 等原有字段保持不变 }; // ... modelConfigs, getInitialState 等保持不变 constgenerateCommand=(values)=>{ const{hardware,modelSize}=values; consthwConfig=modelConfigs[hardware]?.[modelSize]; if(!hwConfig)return`# Error: Unknown hardware/model combination`; let{tp,mem}=hwConfig; constmodelNames={ 'e2b':'google/gemma-4-E2B-it', 'e4b':'google/gemma-4-E4B-it', '12b':'google/gemma-4-12B-it', '31b':'google/gemma-4-31B-it', '26b-a4b':'google/gemma-4-26B-A4B-it', }; // QAT 发布版本保持 bf16 权重 (q4_0-unquantized)，因此唯一变化是模型路径后缀； // TP 和内存需求与标准检查点一致。 constqatSuffix=values.checkpoint==='qat'?'-qat-q4_0-unquantized':''; constmodelPath=`${modelNames[modelSize]}${qatSuffix}`; // ... MTP 相关逻辑保持不变 letcmd=`sglang serve --model-path${modelPath}`; if(tp>1){ cmd+=` \\ --tp${tp}`; } // ... 处理其他选项的命令规则 if(mtpEnabled){ cmd+=` \\ --speculative-algorithm NEXTN`; cmd+=` \\ --speculative-draft-model-path${modelPath}-assistant`; // ... 其余 speculative 参数 } cmd+=` \\ --mem-fraction-static${mem}`; cmd+=` \\ --host 0.0.0.0 --port 30000`; returncmd; }; } } } } } }

评论区精华

该 PR 的 review 过程非常简洁：审核者 zijiexia 直接批准，无讨论或修改要求。唯一的评论来自自动化的 bot 消息（配额提示和文档预览通知），无技术内容。

风险与影响

- 风险：极低。变更仅限于前端组件和文档，无后端逻辑改动。唯一可能的误导是用户误以为 QAT 需要额外参数，但代码注释和文档已明确说明。
- 影响：
  - 用户可通过 UI 直接切换 Standard/QAT 检查点，减少手动拼写错误。
  - 文档覆盖率提升，有助于新用户快速上手。

关联脉络

该 PR 是 Gemma 4 系列文档完善的一部分，与此前的 Gemma 4 部署文档 PR 属于同一功能线。无需跨 PR 关联分析。